DOCUMENT RESUME

ED 177 189 TM 009 632

AUTHOR Corder-Bolz, Charles R.
TITLE An Alternative Model for the Evaluation of Change. Technical Report.
INSTITUTION Southwest Educational Development Lab., Austin, Tex.
SPONS AGENCY Office of Child Development (DHEW), Washington, D.C.
PUB DATE Jun 77
GRANT 90-C-245
NOTE 32p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Gains; *Analysis of Variance; *Evaluation Methods; *Mathematical Models; Research Methodology; Scores; Simulation; Standard Error of Measurement

ABSTRACT

Previous research has indicated that most mathematical models used to evaluate change due to experimental treatment are misleading because the procedures artificially reduced one of the estimates of error variance. Two modified models, based upon the expected values of the variance of scores and difference scores, were developed from a new theoretical perspective. A Monte Carlo study was made involving one-way analysis of variance using the two modified and two basic models. A total of 1,080 experiments were simulated on three treatment levels, six error levels, and three pretest difference levels. The second modified model estimated treatment effects, without being subject to possible distortions due to initial between-group differences, and yielded valid estimates of treatment effects, regardless of the amount of error variance. While the statistic generated by this model has a theoretical weakness, results of this study indicate that it may be a functional solution to the measurement of change controversy. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED177189

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

An Alternative Model for the Evaluation of Change

A Technical Report for the
Television and Social Attitudes Research Grant
Grant No. 90-C-245

by

Charles R. Corder-Bolz, Ph.D.
Southwest Educational Development Laboratory
Austin, Texas

Submitted to

Office of Child Development
Department of Health, Education, and Welfare
Washington, D.C.

June, 1977

TM009 632

Abstract

The results of previous research have indicated that most models used to evaluate change generate misleading appraisals of treatment effects. These invalid assessments were primarily due to the fact that the procedures artifically reduced one of the estimates of the error variance. Two alternative or modified models, based upon the expected values of the variance of scores and difference scores, were developed from a new theoretical perspective. A Monte Carlo study of four models indicates that the second modified model appears to yield reasonably valid estimates of treatment effects. While the statistic generated by the second modified model has a theoretical weakness, the results of this study indicate that it may be a functional solution to the measurement of change controversy.

Most educational and psychological researchers have a serious interest in the assessment of change phenomenon. Much of the research is centered on issues that involve change, such as learning, and development. Most of the procedures developed for the assessment of change are clouded with a great deal of uncertainty. Serious questions have been raised regarding the validity of most of the procedures. Cronbach and Furby (1970) have taken the position that these procedures are so questionable that change itself should not be measured. They suggest, rather, that researchers should design their investigations in such a way that they can measure and observe their constructs without having to assess change.

Certainly, it is difficult for the researcher to compare the relative strengths of the various procedures designed to assess change because they are based upon different assumptions and different perspectives. This lack of common ground has led to different conclusions by different authorities as to how to assess change and to contradictory claims. And, as yet, there is no theoretical basis for resolving these questions regarding how change should be assessed. A Monte Carlo study, which simulated experiments using treatments and control groups which had been pretested and post-tested, evaluated six major models for the assessment of change (Corder-Bolz, in press). In general, all of the six procedures failed to validly estimate the change effects due to the treatments. The one-way analysis of variance of post-test scores, as

expected, was unable to handle the analysis of data in which there were initial or pre-treatment differences between the groups. The two-way analysis of variance using a treatment/control versus pretest/post-test design consistently underestimated the treatment effects. The analysis of covariance of post-test scores, using the pretest scores as the covariate, also substantially overestimated treatment effects when the error variance in the data was less than 60%. Moreover, this procedure proved to be totally unable to handle initial between-group differences and generated totally distorted estimates in such circumstances. The one-way analysis of variance of residualized gain scores procedure proved to generate results very similar to the results generated by the analysis of covariance. The two-way analysis of variance using repeated measures with a treatment/control versus pretest/post-test design resulted in the same arithmetic computations as a one-way analysis of variance of difference scores. These last two procedures were able to handle initial between-group differences, however, they substantially overestimated treatment effects when the error variance in the data was less than 50%.

The basic problem with the more justifiable procedures, i.e., the two-way analysis of variance using repeated measures, one-way analysis of variance using difference scores, lies in the estimates of within-group variance. The statistical structure of these procedures is based upon the generation of two independent estimates of error variance. The variance of the group means provide one estimate of σ_E^2. The variance

within groups provides another and independent estimate of σ_E^2. From these two estimates of variance, a ratio is formed: $F = \frac{MS_B}{MS_F}$. This ratio is distributed as a function of two sampling distributions. A ratio sufficiently in excess of the expected values is interpreted as indicating the presence of treatment effects. The computational structure of the above procedures is based upon the concept of the control of initial differences. Each procedure, in its own way, extracts the pre-test or pre-treatment scores from the post-test scores. In a sense, individuals are used as their own controls. Post-test scores are transformed to a common base or what is sometimes termed "base free." The effect of transforming all individual post-test scores to some kind of measurement scale or base that is free of the individual's pretest differences is to remove the variance among individuals and leave only the differences due to the treatment groups (if there is any treatment effect). Yet, the statistical structure of these models still generates two estimates of error variance: one based upon within-group variance, the other based upon the variance of the group means. Thus, when the within-group variance is reduced or eliminated, by removing individual initial differences, the ratio of the between-group variance over the within-group variance leads to a misleading appraisal of treatment effects. Only experimental error, such as error of measurement or treatment error (individual by treatment interaction) compose the within-group variance. Corder-Bolz (in press) found that an error level below

50% (or test reliabilities greater than 0.707) resulted in highly inflated F ratios. However, data with 50% or more error variance would commonly be regarded as being unacceptable. A researcher interested in investigating change is thus left without a procedure to evaluate the change that might be represented in his data.

The above findings can be represented in a more formal manner. The score of any individual in any treatment group can be thought of as being composed from a linear sum:

$$X = \mu + \alpha_j + \iota_{ij} + e_{ij}$$

This model asserts that the score for observation or individual i in group j is based on the sum of four components: the grand mean of all the J different treatment populations, the effect associated with the particular treatment j, a part which is peculiar to the person i, and another part which is strictly specific to measurement of individual i within treatment group j. The ι_i term is the individual differences component of the particular individual with regard to the particular variable and would be independent of all other terms in the model. The e_{ij} term is the random error associated with the measurement of the particular treatment-individual combination.

For any group j, this model suggests that

(1) $$\bar{X}_j = \Sigma X_{ij}/n_j = \mu + \alpha_j + M_{\iota j} + M_{ej}$$

and that the grand mean

(2) $$\bar{X} = \Sigma_j \Sigma_i X_{ij}/N = \mu + M_\iota + M_e.$$

Therefore the deviation of any treatment group mean from the grand mean can be expressed as

$$(3) \qquad \bar{X}_j - \bar{X} = \alpha_j + M_{i_j} + M_{e_j} - (M_i + M_e).$$

To utilize an analysis of variance model, a sum of squares between groups can be computed and follows that

$$(4) \qquad \text{SS between} = \sum_i n_j (\bar{X}_j - \bar{X})^2 = \sum_i n_j [X_j + M_{i_j} + M_{e_j} - (M_{i_j} + M_e)]^2.$$

In order to determine the statistical implications, the expected value of the SS between can be derived in the following manner.

$$(5) \qquad E(\text{SS between}) = E \sum_j n_j [\alpha_j + M_{i_j} + M_{e_j} - (M_i + M_e)]^2$$

$$(6) \qquad = \sum_j n_j \alpha_j + E \sum_j n_j^2 (M_{i_j} + M_{e_j} - M_i - M_e)^2.$$

On squaring and distributing, the last terms become

$$(7) \qquad \sum_j n_j (M_{i_j} + M_{e_j} - M_i - M_e)^2 = \sum_j n_j M_{i_j}^2 + \sum_j n_j M_{e_j}^2 + \sum_j n_j M_i^2$$

$$+ \sum_j n_j M_e^2 + 2\sum_j n_j M_{i_j} M_{e_j} - 2M_i \sum_j n_j M_{i_j} - 2M_e \sum_j n_j M_{i_j}$$

$$- 2M_i \sum_j n_j M_{e_j} - 2M_e \sum_j n_j M_{e_j} + 2\sum_j n_j M_i M_e.$$

Since

$$(8) \qquad M_e = \sum_j n_j M_{e_j} / N$$

$$(9) \qquad M_i = \sum_j n_j M_{i_j} / N$$

and

(10) $E(\iota_i \cdot e_{ij}) = 0$

the expression of the squaring and distributing simplifies to

(11) $$E\Sigma_j n_j(M_{\iota j} + M_{ej} - M_\iota - M_e)^2 = \Sigma_j n_j E(M_{ej}^2) - NE(M_e^2) + \Sigma_j n_j E(M_{\iota j}^2) - NE(M_i^2)$$

Because

(12) $E(M_{ej}) = E(M_e) = 0$

(13) $E(M_{\iota j}) = E(M_{\iota_j}) = 0$

then for any group j,

(14) $$E(M_e{}^2{}_j) = \sigma^2{}_{M_{ej}} = \frac{\sigma^2{}_e}{n_j}$$

(15) $$E(M_\iota{}^2{}_i) = \sigma^2{}_{M_{\iota j}} = \frac{\sigma^2{}_i}{n_j}$$

Furthermore,

(16) $$E(M_e{}^2) = \sigma^2{}_{M_e} = \frac{\sigma^2{}_e}{N}$$

(17) $$E(M_\iota{}^2) = \sigma^2{}_{M_\iota} = \frac{\sigma^2{}_\iota}{N}$$

Thus

(18) $$E\Sigma_j n_j(M_{\iota j} + M_{ej} - M_\iota - M_e)^2 = \frac{\Sigma_i n_j \sigma^2{}_e}{n_j} - \frac{N\sigma^2{}_e}{N} + \frac{\Sigma_i n_j \sigma^2{}_\iota}{n_j} - \frac{N\sigma^2{}_\iota}{N}$$

(19) $$= (J - 1)\sigma_e{}^2 + (J - 1)\sigma_\iota{}^2$$

By making the appropriate substitutions, the statistical implications are clear

$$(20) \qquad E(\text{SS between}) = \sum_j n_j\sigma^2_j + (J-1)\sigma^2_e + (J-1)\sigma^2_\iota.$$

Customarily, the mean square between is defined as

$$(21) \qquad E(\text{MS between}) = \sigma^2_e + \sigma^2_\iota + \frac{\sum_j n_j\sigma_j^2}{(J-1)}$$

To pursue the analysis of variance implications further, a sum of squares within groups can be represented as follows

$$(22) \qquad X_{ij} - \bar{X}_j = \iota_{ij} + e_{ij} - M_{\iota j} - M_{ej}$$

$$(23) \qquad \text{SS within} = \sum_{ji}\sum(X_{ij} - \bar{X}_j)^2 = \sum_{ji}\sum(\iota_{ij} + e_{ij} - M_{\iota j} - M_{ej})^2$$

In order to determine the statistical implications, the expected value of the SS within can be derived in the following manner.

$$(24) \qquad E(\text{SS within}) = E[\sum_{ji}\sum(X_{ij} - \bar{X}_j)\,]^2 = E[\sum_{ji}\sum(\iota_{ij} + e_{ij} + M_{\iota j} + M_{ej}]^2$$

For any given sample group j,

$$(25) \qquad E[\sum_i(\iota_{ij} - M_{ij})^2/n_j - 1] = \sigma_\iota{}^2$$

$$(26) \qquad E[\sum_i(e_{ij} - M_{ej})^2/n_j - 1] = \sigma_e{}^2$$

and since

$$(27) \qquad E(\iota_{ij} \cdot e_{ij}) = 0$$

then

$$(28) \qquad E(\text{SS within}) = \sum_j \sum_i (X_{ij} - \bar{X}_j)^2$$

$$= \sum_j (n_j - 1)\sigma_t^2 + \sum_j (n_j - 1)\sigma_e^2$$

$$(29) \qquad = (N - J)\sigma_t^2 + (N - J)\sigma_e^2$$

The mean square within can be defined as

$$(30) \qquad MS(\text{within}) = \frac{SS \text{ within}}{(N - J)}$$

and therefore

$$(31) \qquad E(MS \text{ within}) = \sigma_t^2 + \sigma_e^2$$

A statistical test of the differences of the effects of the treatments can be conducted with a ratio of the MS between over the MS within which is distributed as an F with J - 1 and N - J degrees of freedom. This is the traditional model used to evaluate treatment effects and is the basis for commonly used computation procedures, such as the analysis of variance of post-test scores.

The model is appropriate if the data can be reasonably expected to conform to the assumptions that the j group's expected values for error and for individual differences are equal. However, researchers commonly have to deal with conditions or situations in which there is reason to believe that the mean of the individual differences in a group may not

equal the mean of the individual differences in another group. If the subjects are pretested, this non-equivalence will be reflected in pre-treatment mean differences between the groups and could ultimately be confounded with the assessment of the treatment effects. Hence the desire to control for possible pre-treatment group differences.

Expanding upon the above model, the post-test observations can be represented

$$X_{ij_2} = \mu + \alpha_j + \iota_{ij} + e_{ij_2} \tag{32}$$

and the pre-test observations

$$X_{ij_1} = \mu + \iota_{ij} + e_{ij_1} \tag{33}$$

Prior research has found that the analysis of variance of difference scores is able to validly control for pretest differences between group. The difference score can be represented

$$\Delta_{ij} = X_{ij2} - X_{ij1} \tag{34}$$

and with the appropriate substitutions, it is clear that

$$\Delta_{ij} = \alpha_j + e_{ij2} - e_{ij1} \tag{35}$$

Furthermore, it is also clear that

$$\bar{\Delta}_j = \sum_i (X_{ij2} - X_{ij1})/n_j = \alpha_j + M_{ej2} - M_{ej1} \tag{36}$$

$$\bar{\Delta} = \sum_j \sum_i (X_{ij2} - X_{ij1})/N = M_{e_2} - M_{e_1} \tag{37}$$

Therefore, the deviation of any treatment group mean from the grand mean can be expressed as

$$(38) \qquad \bar{\Delta}_j - \bar{\Delta} = \alpha_j + M_{ej2} - M_{ej1} - (M_{e2} - M_{e1})$$

The sum of square between groups can be computed as

$$(39) \qquad \text{SS between} = \sum_j n_j(\bar{\Delta}_j - \bar{\Delta})^2 = \sum_j n_j[\alpha_j + (M_{ej2} - M_{ej1} - M_{e2} + M_{e1})]^2$$

In order to determine the statistical implications, the expected value of the SS between for difference scores can be derived in a manner similar to the above procedure.

$$(40) \qquad E(\text{SS between}) = E\sum_j n_j[\alpha_j + (M_{ej2} - M_{ej1} - M_{e2} + M_{e1})]^2$$

$$(41) \qquad = \sum_j n_j\alpha_j^2 + E\sum_j n_j(M_{ej2} - M_{ej1} - M_{e2} + M_{e1})^2$$

On squaring and distributing, the last term becomes

$$(42) \qquad \sum_j n_j(M_{ej2} - M_{ej1} - M_{e2} + M_{e1})^2 = \sum_j n_j M_{ej2}^2 + \sum_j n_j M_{ej1}^2$$

$$+ \sum_j n_j M_{e2}^2 + \sum_j n_j M_{e1}^2 - 2\sum_j n_j M_{ej2}M_{ej1} = 2M_{ej2}\sum_j n_j M_{ej2}$$

$$+ 2M_{e1}\sum_j n_j M_{ej2} + 2M_{e2}\sum_j n_j M_{ej1} - 2\sum_j n_j M_{e2}M_{e1}$$

$$- 2M_{e1}\sum_j n_j M_{ej1}$$

Since

$$(43) \qquad M_{e_2} = \frac{\sum_j n_j M_{ej_2}}{N}$$

$$(44) \qquad M_e = \frac{\Sigma n_j M_{ej}}{N}$$

and

$$(45) \qquad E(e_{ij_1} \cdot e_{ij_2}) = 0$$

the expression of the squaring and distributing simplifies to

$$(46) \qquad E\Sigma_j n_j(M_{ej2} - M_{ej1} - M_{e2} + M_{e1}) = \Sigma_j n_j E(M_{ej2}^2) - NE(M_{e2}^2) + \Sigma_j n_j E(M_{ej1}^2) - NE(M_{e1}^2)$$

Because

$$(47) \qquad E(M_{ej_2}) = E(M_{e_2}) = 0$$

$$(48) \qquad E(M_{ej_1}) = E(M_{e_1}) = 0$$

then for any group j,

$$(49) \qquad E(M_{ej_2}^2) = \sigma^2 M_{ej_2} = \frac{\sigma^2 e_2}{n_j}$$

$$(50) \qquad E(M_{ej_1}^2) = \sigma^2 M_{ej1} = \frac{\sigma^2 e_1}{n_j}$$

Furthermore,

$$(51) \qquad E(M_{e_2}^2) = \sigma^2 me_2 = \frac{\sigma^2 e_2}{N}$$

$$(52) \qquad E(M_{e_1}^2) = \sigma^2 me_1 = \frac{\sigma^2 e_1}{N}$$

Thus

$$(53) \qquad E\sum_j n_j(M_{ej_2} - M_{ej_1} - M_{e_2} + M_{e_1})$$

$$= \frac{\sum_j n_j\sigma^2_{e_2}}{n_j} - \frac{N\sigma^2_{e_2}}{N} + \frac{\sum_j n_j\sigma^2_{e_1}}{n_j} - \frac{N\sigma^2_{e_1}}{N}$$

$$(54) \qquad = (J - 1)\sigma^2_{e_2} + (J - 1)\sigma^2_{e_1}$$

By making the appropriate substitutions, the statistical implications are evident

$$(55) \qquad E(\text{SS between}) = \sum_j n_j\alpha_j^2 + (J - 1)\sigma^2_{e_2} + (J - 1)\sigma^2_{e_1}$$

The mean square between for difference scores can be defined as

$$(56) \qquad \text{MS between} = \frac{\text{SS between}}{J - 1}$$

and therefore

$$(57) \qquad E(\text{MS between}) = \sigma^2_{e_2} + \sigma^2_{e_1} + \frac{\sum_j n_j\alpha_j^2}{(J - 1)}$$

To puruse the implications of an analysis of variance of difference scores further, a sum of squares within groups can be represented as follows

$$(58) \qquad \Delta_{ij} - \bar{\Delta}_j = (X_{ij_2} - X_{ij_1}) - \left(\frac{\sum_j (X_{ij_2} - X_{ij_1})}{n_j}\right)$$

$$(59) \qquad = e_{ij_2} - e_{ij_1} - (M_{ej_2} - M_{ej_1})$$

$$(60) \qquad \text{SS within} = \sum_{ji}\sum(\Delta_{ij} - \bar{\Delta}_j)^2 = \sum_{ji}\sum[(e_{ij_2} - e_{ij_1}) - (M_{ej_2} - M_{ej_1})]^2$$

The expected value of SS within for difference scores can be derived in the following manner.

$$(61) \qquad E(SS\ within) = E[\sum_{ji}\sum(\Delta_{ij} - \bar{\Delta}_j)^2] = E[\sum_{ji}\sum(e_{ij_2} - e_{ij_2} - M_{ej_2} + M_{ej_1})^2]$$

For any given sample group j,

$$(62) \qquad E\left[\frac{\Sigma(e_{ij_2} - M_{ej_2})^2}{n_j - 1}\right] = \sigma^2_{e_2}$$

$$(63) \qquad E\left[\frac{\Sigma(e_{ij_1} - M_{ej_1})^2}{n_j - 1}\right] = \sigma_{e_1}$$

and since

$$(64) \qquad E(e_{ij_2} \cdot e_{ij_1}) = 0$$

and if

$$(65) \qquad Z = c_1Y_1 + c_2Y_2 + \ldots + C_nY_n$$

then

$$(66) \qquad \sigma_z{}^2 = C_1{}^2\sigma^2_{Y_1} + C_2{}^2\sigma^2_{Y_2} + \ldots + C_n{}^2\sigma^2_{Y_n}$$

$$(67) \qquad E\ (SS\ within) = \sum_j E\sum_i(\Delta_{ij} - \bar{\Delta}_j)^2$$

$$(68) \qquad = \sum_j(n_j - 1)\sigma^2_{e_2} + \sum_j(n_j - 1)\sigma^2_{e_1}$$

$$(59) \qquad = (N - J)\sigma^2_{e_2} + (N - J)\sigma^2_{e_1}$$

The mean square within for differences scores can be defined as

$$(70) \qquad MS(within) = \frac{SS(within)}{(N - J)}$$

and therefore

$$(71) \qquad E(\text{MS within}) = \sigma^2_{e_2} + \sigma^2_{e_1}$$

A statistical test of the differences of the effects of the treatments can be conducted with a ratio of the MS between over the MS within which is distributed as an F with J - 1 and N - J degrees of freedom.

When the analysis of variance of post-test scores and the analysis of variance of difference scores, it is evident as to why the use of difference scores has proven to be an effective approach to the assessment of treatment effects when pre-treatment differences between the groups are possible. However, it can also be seen why the analysis of variance of difference scores yields highly inflated evaluations of the treatment effects when the experimental error, such as measurement unreliability, is low (or less than 50%). The analysis of variance of post-test scores approach evaluates the treatment effects within the context of error variance and individual differences variance.

$$(72) \qquad E(\text{F post-test scores}) = \frac{\sum n_j \alpha_j^2 + \sigma_i^2 + \sigma_e^2}{\sigma_i^2 + \sigma_e^2}$$

The analysis of variance of difference scores approach evaluates the treatment effects only within the context of experimental error variance.

$$(73) \qquad E(\text{F difference scores}) = \frac{\sum n_j \alpha_j^2 + \sigma^2_{e_2} + \sigma^2_{e_1}}{\sigma^2_{e_1} + \sigma^2_{e_1}}$$

At stake is a broader theoretical issue. The ultimate purpose of such analyses is to make inferences with regard to the treatments' effects upon a population of individuals. Therefore, it would appear to be more legitimate to conduct an assessment of the impact of a treatment in terms of a population and a population of measurement error rather than assessment only in terms of one (or two) populations of measurement error. Even though it may be desirable to correct for possible pre-treatment differences between groups, and even though it is statistically legitimate to analyze difference scores in terms of experimental error variance, the resultant statistic has been demonstrated to lead to erroneous inferences of the effects of treatments upon individuals.

As a solution to this Scylla and Charybdian dilemma, it may be possible to form a "hybrid" F-ratio. The presence of pre-treatment differences can be represented in pretest scores as

$$X_{ij_1} = \mu + \pi_j + \iota_{ij} + e_{ij_1} \tag{74}$$

and in post-test scores as

$$X_{ij_2} = \mu + \pi_j + \alpha_j + \iota_{ij} + e_{ij_2} \tag{75}$$

where

$$\Delta_{ij} = \alpha_j + e_{ij_1} - e_{ij_2} \tag{76}$$

with π_j representing the mean pre-treatment differences between group. The expected value of the F-ratio of an analysis of post-test scores

where π_j is non-zero would be of the form

$$(77) \qquad E(F \text{ post-test scores}) = \frac{\sum n_j \alpha_j^2 + \sum n_j \pi_j^2 + \alpha_\iota^2 + \sigma^2_{e_2}}{\sigma_\iota^2 + \sigma^2_{e_2}}$$

Similarly, the expected value of the F-ratio of an analysis of variance of difference scores were π_j is non-zero would be of the form

$$(78) \qquad E(F \text{ difference scores}) = \frac{\sum n_j \alpha_j^2 + \sigma^2_{e_2} + \sigma^2_{e_1}}{\sigma^2_{e_2} + \sigma^2_{e_1}}$$

It would be desirable to be able to generate an F-ratio where π_j is non-zero which would be of the form

$$(79) \qquad E(F_?) = \frac{\sum n_j \alpha_2^2 + \sigma_\iota^2 + \sigma_e^2}{\sigma_\iota^2 + \sigma_e^2}$$

An appropriate estimate of the denominator of this desired statistic is available from the MS within for the analysis of variance of post-test scores

$$(80) \qquad E(\text{MS within}) = \sigma_\iota^2 + \sigma^2_{e_2}$$

The numerator could be approximated by the MS between for the analysis of variance of difference scores and would result in a modified F-ratio of the form

$$(81) \qquad E(F_m{}^1) = \frac{\sum_j n_j \sigma_e{}^2 + \sigma^2{}_{e_2} + \sigma^2{}_{e_1}}{\sigma^2{}_{e_2} + \sigma_\iota{}^2}$$

When the experimental error variance is low, this modification would reduce the highly inflated values generated by the F-ratios based upon the difference score analysis and would eliminate the distortion due to pre-treatment differences. However, this clearly would not provide a completely valid evaluation of the treatment effects.

A second alternative could also estimate the denominator with the MS within for the analysis of variance of post-test scores. The numerator could be estimated from a pseudo mean square.

$$(82) \qquad \text{MS bet. mod. 2} = \text{MS bet (diff scores)} - \text{MS within (diff scores)} + \text{MS within (post-test scores)}$$

$$(83) \qquad E(\text{MS bet. mod. 2}) = (\sum_j n_j \alpha_j{}^2 \quad \sigma^2{}_{e_1} + \sigma^2{}_{e_2}) - (\alpha^2{}_{e_1} + \alpha^2{}_{e_2}) + (\alpha_\iota{}^2 + \sigma^2{}_{e_2})$$

This second modified F-ratio would be of the form

$$(84) \qquad E(F_{m_2}) = \frac{\sum_j n_j \alpha_j{}^2 + \sigma_\iota{}^2 + \sigma^2{}_{e_2}}{\sigma^2{}_\iota + \sigma^2{}_{e_2}}$$

This second modified F-ratio would resolve the previous problems. However, estimates of the experimental error variance and the individual

differences for the numerator and the denominator are not independent estimates, and therefore, violate a basic assumption of the general form of the F-ratio statistic. Since there appears to be no completely legitimate solution from a theoretical perspective to this important problem, the question is thus raised as to whether or not either of the modified F-ratios represent an adequate functional solution to the problem of analyzing change.

Method

The relative strengths of the analysis of variance of post-test scores, analysis of variance of difference scores, the first modified analysis of variance of difference scores and the second modified analysis of variance of difference scores for evaluating change were empirically tested. The approach was to simulate experiments utilizing a treatment versus control-group design in which the hypothetical subjects were pre-tested and post-tested. Using a random number generator, populations of hypothetical subjects were generated such that the populations conformed to specific parameter values. Then these populations were sampled in a manner similar to the way a two-group experiment would be conducted. The population parameters were established in such a way that the effects of treatment, error, and pre-treatment differences were simulated. Populations of 6,500 subjects were generated. From the pretest control group population, samples of 20 observations were randomly selected. Likewise, from the pretest treatment group population, 20 observations were randomly

selected. The two populations from which the two respective samples were randomly drawn, had the same mean and variance (except as modified by simulated pretest differences). The post-test control group scores were the same as the pretest control group scores except as modified by the levels of simulated error variance.

The simulation of treatment effects was accomplished by adding a constant to the pretest treatment scores to generate the post-test treatment scores. The value of the constant depended upon the treatment condition being simulated. Three such treatment conditions were simulated. In the first condition, the population of post-test control group scores and the population of post-test treatment group scores had identical values for the mean. In the second treatment condition simulated, the mean of the post-test control group population differed such that the expected value of the z score of the difference between sample means that would have an associated probability of 5%. A third treatment condition was simulated by generating a population of post-test control group scores and a population of post-test treatment group scores with a difference between the two respective population means such that the expected value of the z score of the difference of sample means would have an associated probability of 1%.

The particular levels of error variance were simulated by manipulating the relative proportion of variance of the true scores and variance of the error components. Six levels of error variance were

simulated: 0%, 10%, 25%, 50%, and 60%. Error was simulated by adding an error component to the pretest and post-test scores. The value of the error component was random and independent of the value of the pretest and post-test scores. In such a manner, observed scores were generated that were composed of true scores plus an error component.

In order not to confound the magnitude of error variance with the magnitude of observed score variance, the variance of observed scores σ_Y^2 was maintained at a constant 1.0. The observed scores were a linear combination of true scores and error components which were respectively multiplied by their weights C_1 and C_2. If X_1 is from a population of true scores, and if X_2 is from a population of error terms, then the proportion of error variance is observed scores Y_{ij}, which are a linear combination of X_1 and X_2, can be determined by the weights of the linear combination.

If $Y_{ij} = C_1X_1 + C_2X_2$ or if $\mu_Y = C_{1\mu X_1} + C_{2\mu X_2}$,

Then $\sigma_Y^2 = C_1^2 + \sigma_{X_1}^2 + C_2^2 \cdot \sigma_{X_2}^2$.

If $\sigma_{X_1}^2 = 1$ and $\sigma_{X_1}^2 = 1$,

Then $\sigma_Y^2 = C_1^2 + C_2^2$.

Thus if $\sigma_Y^2 = 1$,

Then $C_1^2 + C_2^2 = 1$.

From the above equations, it can be seen that C_1^2 plus C_2^2 should equal 1.0 in all simulated error conditions. The proportion of error variance, σ_e^2, was therefore equal to the square of the linear weight, C_2. Thus, the weights for the linear combination could be computed by taking the square root of the respective percentages of true score variance and error variance.

Three levels of initial between-group differences were simulated. In the first level, the pretest control group population and the pretest treatment group population had the same mean and variance. This resulted in a situation in which the differences between the populations of the two respective means were such that the expected value of the z score of sampled means was 0. In the second level of simulated initial between-group differences, the population of pretest scores and the population of post-test scores were such that the difference between the two respective population means resulted in the expected value of the z score for the difference of sample means having a value of 1.96. The probability associated with such a z score value is 0.05. The third level of the simulated initial between-group differences used a pretest control group population and a pretest treatment group population such that the difference between the means of the two respective populations resulted in a z score of the difference between sample means having a value of 2.58. A z score having such a value of 2.58 has an associated probability of 0.01.

In summary, populations were generated and subsequently sampled which met the following definitions:

Pretest control: $Y_{i11} = C_1T_{i11} + C_2E_{i11}$

Pretest treatment group: $Y_{i21} = C_1T_{i21} + \pi + C_2E_{i21}$

Post-test control: $Y_{i12} = C_1T_{i11} + C_2E_{i12}$

Post-test treatment group: $Y_{i22} = C_1T_{i21} + \alpha + \pi + C_2E_{i22}$

where T is a true score from a standard normal population, $T \sim N(1,0)$, E is the error component from a standard normal population, $E \sim N(1,0)$, α is the treatment effect, π is the initial between-group difference, C_1 is the weight for the true scores, and C_2 is the weight for the error components.

In each case, a population of 6,500 was generated. From these populations, samples of 20 pretest control group scores, 20 pretest treatment group scores, 20 post-test control group scores, and 20 post-test treatment group scores, were randomly selected. Each data set, in effect, represented a simulated experiment. The three treatment levels, six error levels, and three pretest difference levels resulted in 54 experiments being simulated. Each of these 54 original experiments was replicated 50 times. Thus, a total of 1,080 experiments was simulated.

Each data set was then analyzed with the one-way analysis of variance of post-test scores model, the one-way analysis of variance of difference scores model and the first modified analysis of variance of difference scores model and the second modified analysis of variance of difference

scores model. Each model produced an estimate of the change caused by the simulated treatment. This estimate was in the form of an F ratio. The mean F ratio over the 50 replications was computed for each of the 54 simulated conditions for the models. The mean F ratios generated by the first and second modified analysis of variance of difference scores models were then compared to the expected F ratios, and the F ratios generated by the analysis of variance of difference scores model and the analysis of variance of post-test scores model.

Results

The first and second modified analysis of variance of difference scores proved to be computationally easy. An example summary table is presented in Table 1.

The mean F ratios as presented in Table 2 for the one-way analysis of variance of post-test scores in the simulated experiments in which there were no pretest differences are equivalent to the expected values. This indicates that the random number generator was functioning properly. There is a slight tendency, though, for the observed mean value to be larger than the expected value, probably due to the highly skewed nature of the F distribution.

The mean F ratios for the one-way analysis of variance of difference indicated two trends. This model in the condition of 50% error generated mean F ratios of 0.749, 5.785, and 6.688 when there were no pretest differences in the population and mean F ratios of 1.057, 6.881, and 7.084

when there were substantial pretest differences. The expected values were respectively 0.464, 4.098, and 7.353. However, in the condition of 10% error, this model generated mean F ratios of 1.459, 21.423, and 35.404 when there were no pretest differences and mean F ratios of 0.989, 19.248, and 37.732 when there were substantial pretest differences. Again, the expected values were respectively 0.464, 4.098, and 7.353.

The first modified model generated mean F ratios which generally were equivalent to the expected values. For example, in the 35% error condition at the second level of treatment, the mean F ratios for the first, second, and third levels of pretest differences respectively were 3.707, 4.199, and 4.008. However, when the error level was low, this model apparently tended to generate mean F ratios lower than the expected value, while it appeared to generate mean F ratios slightly higher than the expected value when the error level was high.

The second modified model appeared to generate F ratio more equivalent to the expected value. For example, at the third level of pretest differences and the second level of treatment effects, the mean F ratios for the six levels of error were respectively 4.910, 4.076, 5.041, 4.221, 6.520, and 4.317, where the expected value was 4.098.

Discussion

The weakness of the traditional analysis of variance of difference scores is evident. When the error level is less than 50%, this model clearly yields invalid estimates of treatment effects.

The first modified model appears to be more robust than the traditional model. However, its estimates of the treatment effects appear to be slightly influenced by the amount of error. Looking back at the expected values of the mean squares between and within, the reason for this bias can be seen. For the first modified model

$$E(\text{MS between}) = \Sigma_j n_j \alpha^2 + \sigma_{e_1}^2 + \sigma^2_{e_2}$$

$$E(\text{MS within}) = \sigma_{\iota}^2 + \sigma^2_{e_2}$$

The bias stems from the fact that $\sigma_{e_1}^2$ does not always equal σ_{ι}^2. When error variance is less than 50%, the generated F ratio will tend to be slightly lower than if $\sigma_{e_1}^2$ did equal σ_{ι}^2. Likewise, the generated F ratio will tend to be slightly higher when the error variance is greater than 50%.

The second modified model appears to be robust with respect to the amount of pre-treatment differences between groups and with respect to the proportion of error variance. While it is theoretically bothersome that this model generates a statistic based upon a redundant use of an estimate of the error variance, it does appear to generate reasonably valid estimates of treatment effects. The mean values of the computed F ratios proved to be close to the expected values of the F ratios. The differences between the observed and the expected values are reasonably within sampling variance. Importantly, these data indicate that the second modified model estimates treatment effects without being subject

to possible distortions due to initial between-group differences. In light of the previous research which has shown that most procedures are very sensitive to the amount of error variance in the data, it is also significant that the modified model yields valid estimates of treatment effects regardless of the amount of error variance. The results thus appear to support the theoretical foundation of this model.

Summary

The results of previous research have indicated that most models used to evaluate change generate misleading appraisals of treatment effects. These invalid assessments were primarily due to the fact that the procedures artificially reduced one of the estimates of the error variance. Two alternative or modified models, based upon the expected values of the variance of scores and difference scores, were developed from a new theoretical perspective. A Monte Carlo study of four models indicates that the second modified model appears to yield reasonably valid estimates of treatment effects. While the theoretical structure of the statistic generated by the second modified model has a weakness, the results of this study indicate that it may be a functional solution to the measurement of change controversy.

References

Corder-Bolz, C. R. We can measure change, but can we evaluate it? Educational and Psychological Measurement, in press.

Cronbach, L. J., and Furby, L. How we should measure "change"--or should we? Psychological Bulletin, 1970, 74, 68-80.

Table 1

Summary Table of Analyses of an Example Data Set*

One-way Analysis of Variance of Post-test Scores

	Sum of Squares	df	Mean Square	F ratio
Between	5.077	1	5.077	4.726
Within	40.317	38	1.074	
Total	45.394	39		

One-way Analysis of Variance of Difference Scores

	Sum of Squares	df	Mean Square	F ratio
Between	4.566	1	4.566	21.443
Within	8.091	38	0.213	
Total	12.657	39		

First Modified One-way Analysis of Variance of Difference Scores

	Sum of Squares	df	Mean Square	F ratio
Between	4.566	1	4.566	4.250
Within	40.317	38	1.074	
Total	unknown	39		

Second Modified One-way Analysis of Variance of Difference Scores

	Sum of Squares	df	Mean Square	F ratio
Between	5.427	1	5.427	5.053
Within	40.917	38	1.074	
Total	unknown	39		

*data with 10% error and no initial between-group differences

Table 2

Mean F ratios Generated by (Traditional) Analysis of Variance of Difference Scores and Modified Analysis of Variance of Difference Scores

Treatment Level	Expected Value	Post-Test Difference											
		Level 1				Level 2				Level 3			
		One-Way ANOVA of Post-Test	Traditional Difference Score ANOVA Model	First Modified Difference Score ANOVA Model	Second Modified Difference Score ANOVA Model	One-Way ANOVA of Post-Test	Traditional Difference Score ANOVA Model	First Modified Difference Score ANOVA Model	Second Modified Difference Score ANOVA Model	One-Way ANOVA of Post-Test	Traditional Difference Score ANOVA Model	First Modified Difference Score ANOVA Model	Second Modified Difference Score ANOVA Model
		0% Error											
1	0.464	1.024	0	0	0.999	4.127	0	0	0.999	7.282	0	0	0.999
2	4.038	4.323	-	3.985	4.985	18.413	-	3.148	4.149	21.161	-	3.910	4.910
3	7.353	7.343	-	6.033	7.032	23.853	-	7.458	8.452	28.808	-	7.057	7.557
		10% Error											
1	0.461	1.094	1.459	0.293	1.075	4.467	1.340	0.258	1.059	6.897	0.989	0.186	0.980
2	4.038	3.794	21.423	4.318	5.106	17.996	20.953	3.932	4.795	24.375	19.248	3.391	4.076
3	7.353	7.336	35.404	7.115	7.397	21.338	39.294	7.282	7.465	28.936	37.732	7.463	8.256
		25% Error											
1	0.464	1.254	0.985	0.549	0.996	4.968	0.813	0.333	0.904	8.321	0.986	0.477	0.944
2	4.038	4.451	9.095	4.669	5.147	17.376	10.005	3.839	4.306	24.251	9.156	4.554	5.041
3	7.353	8.05.	15.922	7.071	7.545	23.594	15.912	6.738	7.219	28.827	15.211	6.832	7.295
		35% Error											
1	0.464	1.022	0.918	0.692	0.923	4.751	0.558	0.412	0.679	7.819	1.231	0.909	1.141
2	4.038	4.372	4.978	3.707	3.990	16.499	6.027	4.193	4.461	22.295	6.650	4.008	4.221
3	7.353	7.653	11.407	7.570	7.881	23.514	10.381	7.693	7.964	28.432	11.929	7.283	7.550
		50% Error											
1	0.464	0.822	0.749	0.668	0.742	4.336	1.349	1.390	1.371	7.479	1.057	1.024	1.048
2	4.038	5.776	5.785	5.983	5.932	18.849	5.623	5.903	5.826	26.417	6.891	6.523	6.520
3	7.353	7.985	6.688	6.989	6.902	22.348	8.417	8.447	8.379	27.773	7.084	7.073	7.064
		60% Error											
1	0.464	1.025	1.139	1.363	1.209	4.479	0.681	0.806	0.730	8.012	0.943	0.963	0.868
2	4.038	4.312	3.979	5.012	4.779	15.410	4.683	5.385	5.226	22.707	3.758	4.549	4.317
3	7.353	7.311	7.968	7.975	7.464	22.743	8.584	7.761	7.482	28.614	6.541	7.687	7.457